Press Release

2026 年 11 月 11 日

東京科学大学

横浜市立大学

東北大学

神奈川県立産業技術総合研究所

言語モデルを用いた材料物性の予測を軽量化

－材料情報の文章入力が可能で、汎用的な PC でも物性予測を可能に－

【ポイント】

○材料の情報を文章で表すことで、専門的な数値指標を設計せずに材料物性を予測する軽量な手法「ZEBRA-Prop」を開発

○従来の LLM を用いた手法より学習時間を約 95%短縮、PC でも学習が可能に

○合成条件や実験記録など、文章で記述された情報の活用への展開に期待

【概要】

東京科学大学（Science Tokyo）物質理工学院 材料系の山本竜馬修士課程学生、総合研究院 フロンティア材料研究所の高橋亮准教授、大場史康教授（神奈川県立産業技術総合研究所（KISTEC）プロジェクトリーダー兼任）は、横浜市立大学 大学院生命医科学研究科の寺山慧研究教授、東北大学 金属材料研究所の熊谷悠教授らと共同で、**大規模言語モデル**（**LLM**、用語 1）を再学習することなく、材料物性を予測できる軽量な機械学習手法「ZEBRA-Prop（ゼブラプロップ）」を開発しました。

新規材料の探索では、材料の構造や組成から性質を予測する機械学習手法が広く使われています。こうした手法では、実験やシミュレーションにより物性値が分かっている多数の材料のデータを予測モデルにあらかじめ学習させておき、その学習済みのモデルを用いて新しい材料の物性を高速に予測します。近年、材料の情報を自然言語の文章で表現し LLM で予測する手法が登場しました。しかし、この手法は予測したい材料物性ごとに LLM 自体を**ファインチューニング**（**再学習**、用語 2）する必要があり、高性能・大規模な計算資源を持たない研究者が使うことは容易ではありませんでした。

研究チームは、モデルの学習の際に LLM を再学習せず、後段の小規模なニューラルネットワークだけを学習する方式を考案しました。これにより、学習時間を従来の LLM を用いた手法より約 95%短縮できました（約 2,200 件のデータでの比較）。さらに、材料の情報を複数の短い文章に分けて表現し、重みを学習しながら統合することで、LLM が一度に読める文章量の制約も緩和し、精度も従来の LLM を用いた手法に近い水準を保っています。約 2,200 件の材料データであれば、高性能な **GPU**（用語 3）を使わずに、一般的な PC を用いて数十分程度で学習できることも確認しています。

本手法は、自然言語の文章で記述できる情報であれば入力に使えるため、合成条件や実験記録など、文章で記述された実験データから材料物性を予測する技術への展開が期待されます。

本成果は、11 月 11 日付（現地時間）の「*Advanced Intelligent Systems*」誌に掲載されました。

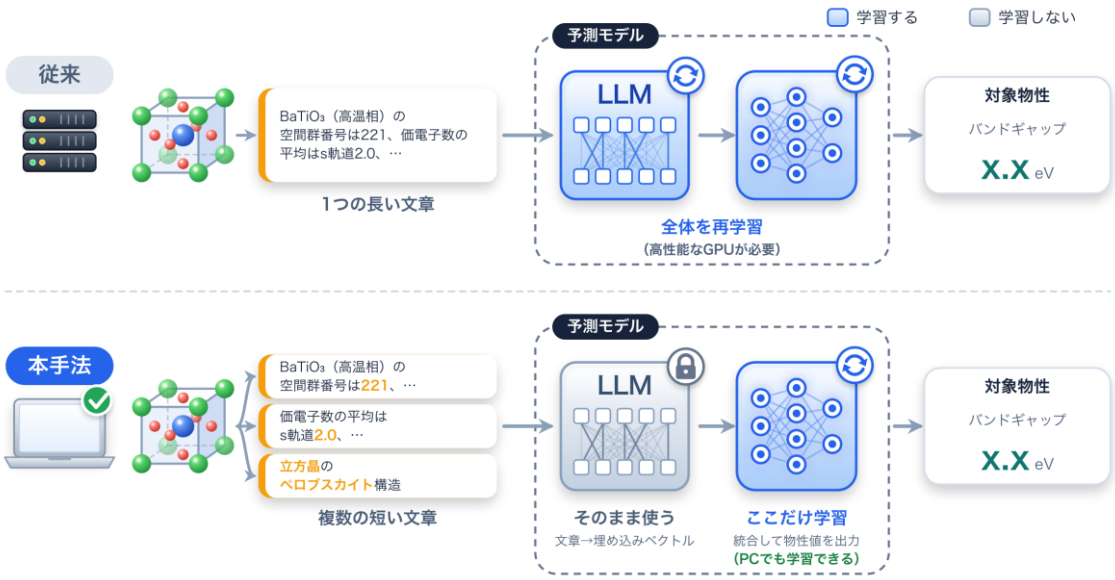


図 1. ZEBRA-Prop の学習過程の概略図。点線内が予測モデルであり、あらかじめ物性値が既知である大量のデータを学習することで構築される。従来手法（上）は LLM を含む全体を学習するが、本手法（下）は後段の小規模なニューラルネットワークだけを学習するため、学習が格段に高速化されている。ここではバンドギャップを対象物性とし、$BaTiO_3$（チタン酸バリウム）の高温相を例として示している。

●背景

新しい材料の開発では、膨大な候補物質の中から目的の性質を持つものを絞り込むことが重要です。そのため、材料の構造から電子の性質や安定性を予測する機械学習が広く使われています。例えば、理論計算で求めた材料物性のデータを用いて、原子配列と材料物性の関係を学習する手法が、高い精度を示してきました。その代表が、原子のつながり（原子間距離や角度の情報）をネットワーク（グラフ）として扱う**グラフニューラルネットワーク**（**GNN**、用語 4）です。しかし、GNN は原子配列が明確に定まっていることが前提です。例えば実験で得られたデータを活用しようとする際に、合成条件や測定条件のような情報など、原子配列として表現できないため GNN の入力として活用することが難しい場合が少なくありませんでした。

その一方で、例えば原子配列の情報を「バリウム原子は 12 個の酸素原子に囲まれている」といった文章で表し、大規模言語モデル（LLM）で材料物性を予測する手法（LLM-Prop）［参考文献 1］が近年提案されました。一般的な機械学習では、材料の情報を、

すべての材料で共通の項目からなる数値の組（数値指標）に直して入力する必要があり、どの情報をどう数値で表すかの設計に専門知識が必要でした。文章であれば、「空間群番号は 221」のような数値も、「立方晶」のような言葉の情報も、さらに合成条件や測定条件といった実験の情報もそのまま書き込めるため、専門外の研究者にも使いやすいと期待されています。ただし、この手法は予測したい材料物性ごとに巨大な LLM 全体を再学習する必要があり、そのためには高性能・高価な計算機が必要でした。また、LLM が一度に処理できる文章量には上限があり、長い文章は途中までしか読み込めないという課題もありました。そこで本研究では、実験研究者でも手元の PC で使える、より軽量な手法の開発に取り組みました。

●研究成果

ZEBRA-Prop（Zero-Shot Embedding-Based Rapid and Accessible Regression Model for Materials Properties：ゼロショット埋め込みに基づく、高速で使いやすい材料物性回帰モデル）は、次の 4 段階で材料物性を予測します（図 1）。まず材料の情報を、観点の異なる複数の短い文章で表現します。次に、各文章を材料科学に特化した LLM で**埋め込みベクトル**（用語 5）に変換します。続いて、得られたベクトルを学習可能な重みで足し合わせて 1 つにまとめ、最後に小規模なニューラルネットワークに入力して材料物性の値を出力します。

本研究の性能評価では、入力する文章として、元素の性質や対称性などを表す既存の数値指標（記述子）を文章の形にしたもの（10 種類）と、結晶構造を自動で文章にするツールが生成した自然な文章（2 種類）の、合計 12 文を用いました。入力文章を短く分けることで、LLM が一度に読める文章量の制約を緩和しながら、多様な情報を取り込めます。さらに、化学式の括弧を外して元素と数だけで表す（例：$Cu(NO_3)_2$を Cu 1 N 2 O 6 と表記）、数値を一定倍して整数に丸める、といった前処理を加え、LLM が扱いやすい形に整えることで予測精度が向上することを確認しました。

性能評価には、**第一原理計算**（用語 6）で求めた約 2,200 件の材料データと、材料データベース **Materials Project**（用語 7）に由来する約 14 万件のデータを用いました。いずれも、データの一部で予測モデルを学習させ、学習に使っていない残りの材料について、予測値を第一原理計算の値と比べて精度を確かめました。約 2,200 件のデータで高性能な GPU1 基を使って学習時間を比較したところ、本手法の学習時間は従来の LLM を用いた手法（LLM-Prop）より約 95%短く、従来使われてきた GNN（**CGCNN**［参考文献 2］・**ALIGNN**［参考文献 3］、用語 8）と比較しても短い時間で済みました（図 2）。また学習時間の約 8 割は、LLM による埋め込みベクトルへの変換が占めますが、この変換は一度行えば、別の予測対象となる材料物性の学習にも使い回すことができます。

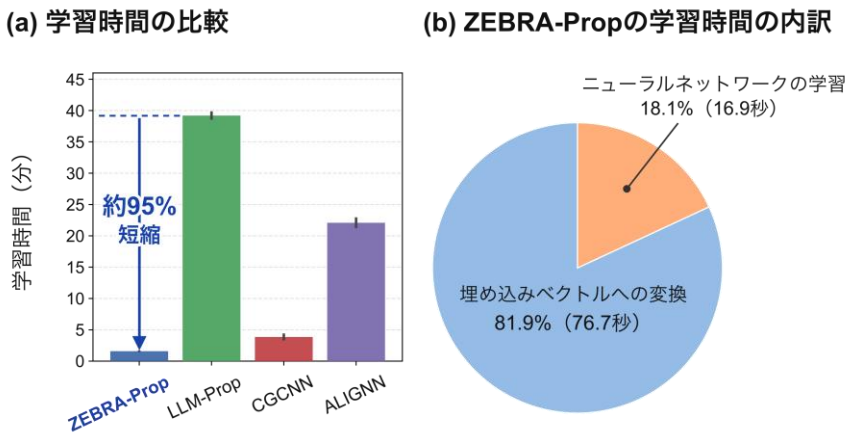


図 2. 学習時間の比較。（a）ZEBRA-Prop の学習時間は、LLM-Prop より約 95%短く、従来の GNN（CGCNN・ALIGNN）よりも短い。（b）ZEBRA-Prop の学習時間の約 8 割は、埋め込みベクトルへの変換が占める（約 2,200 件のデータ、GPU1 基、5 分割交差検証の平均）。

これは GPU などの高価な計算資源を使わず一般的な PC（Apple M2 チップ、メモリ 16 GB）でも、約 2,200 件のデータなら、埋め込みベクトルへの変換に数十分、材料物性予測モデルの学習に数分で実行できるほどの高速化です。

約 2,200 件のデータを学習させた予測モデルの精度は、**バンドギャップ**（用語 9）、形成エネルギー（物質の安定性の指標）、誘電率のいずれも、高い精度が報告されている LLM-Prop に近い水準でした。バンドギャップと形成エネルギーでは、複数の文章を統合すると 1 種類の文章だけの場合より精度が高まりました。材料科学に特化した LLM の利用や、化学式・数値の前処理も精度向上に寄与しました。一方で原子配列を直接扱う GNN に精度では及ばず、約 14 万件のデータでは LLM-Prop との性能差も若干大きくなりますが、少ない計算資源と実用的な精度の両立を図った点が本手法の特長です。また、入力は文章で記述できる情報であればよいため、記述子の設計や原子配列の特定が難しい場合や、合成条件・測定条件といった実験の情報への応用も期待されます。

●社会的インパクト

材料の研究現場では、実験を行う研究者が貴重なデータを蓄積している一方で、機械学習を導入するには物質の情報を数値化するための専門知識と高性能な計算機が必要でした。ZEBRA-Prop は、文章という身近な形で情報を入力でき、手元の PC で実行できるほど高速なため、導入の負担を大きく減らせます。実験研究者が素早く予測モデルを作れるようになれば、有望な候補の絞り込みが進み、実験とデータ科学をつなぐ新材料の開発が加速することが期待されます。また、本手法のプログラムと評価に用いたデータセットは公開しています（https://github.com/oba-group/ZEBRA-Prop、https://github.com/oba-group/ZEBRA-dataset）。

●今後の展開

実験データには、合成条件、熱処理の履歴、測定条件、観察メモなど、文章で記述された情報が含まれます。ZEBRA-Prop は文章を入力とするため、こうした実験条件や作製プロセスの記述から、実験で測定される材料物性を直接予測する方向への展開が期待されます。今後は、LLM の高性能化や入力する文章の質と多様性の向上による精度の改善を試みます。さらに試料の写真や顕微鏡像など、他の種類のデータも併用できるマルチモーダルな機械学習モデルとの組み合わせも視野に入れています。

●付記

本研究は、日本学術振興会 科学研究費助成事業（JP24K08562、JP24H00376）、文部科学省 データ創出・活用型マテリアル研究開発プロジェクト事業（JPMXP1122683430）、国際・産学連携インヴァースイノベーション材料創出プロジェクト（DEJI²MA）、および神奈川県立産業技術総合研究所（KISTEC）脱炭素化対策事業の助成を受けて行われました。計算には東京科学大学のスーパーコンピュータ TSUBAME4.0 を使用しました。

【参考文献】

[1] A. Niyongabo Rubungo, C. Arnold, B. P. Rand, and A. B. Dieng, "LLM-Prop: predicting the properties of crystalline materials using large language models", *npj Computational Materials*, 2025, 11, 186.

[2] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties", *Physical Review Letters*, 2018, 120, 145301.

[3] K. Choudhary and B. DeCost, "Atomistic Line Graph Neural Network for improved materials property predictions", *npj Computational Materials*, 2021, 7, 185.

【用語説明】

（1） **大規模言語モデル（LLM）**：大量の文章を学習し、文章の理解や生成ができる機械学習モデル。ChatGPT などが代表例。

（2） **ファインチューニング（再学習）**：学習済みのモデルを、特定の課題に合わせて追加で学習し、内部の数値を調整すること。精度は上がりやすいが、多くの計算資源と時間を要する。

（3） **GPU**：もともとは画像処理用に開発された演算装置で、機械学習のような大量の計算を高速に実行できる。高価なため、利用できる環境が限られる。

（4） **グラフニューラルネットワーク（GNN）**：原子を点、原子間の結合を線とするグラフ構造を入力として、材料物性を予測する深層学習の手法。原子配列が分

かっている材料で高い精度を示す。

(5) **埋め込みベクトル**：文章の意味を、多数の数値で表したもの。意味が近い文章ほど、近い数値になり、コンピュータが文章の意味を扱えるようになる。

(6) **第一原理計算**：量子力学の基本原理に基づく理論計算。物質の電子構造やエネルギーを求め、さまざまな材料物性を予測できる。

(7) **Materials Project**：第一原理計算で得られた物質の構造や物性に関する情報を提供する、材料科学分野の大規模オープンデータベース。

(8) **CGCNN・ALIGNN**：GNN の代表的な手法。CGCNN は原子間の結合関係を、ALIGNN は結合の長さに加えて結合間の角度も考慮して、材料物性を予測する。

(9) **バンドギャップ**：物質中の電子が取りうるエネルギーの間にある隙間の大きさ。電気の流れやすさや光の吸収を左右する、半導体の重要な指標。

【論文情報】

掲載誌：*Advanced Intelligent Systems*
論文タイトル：ZEBRA-Prop: A Zero-Shot Embedding-Based Rapid and Accessible Regression Model for Materials Properties（ZEBRA-Prop：ゼロショット埋め込みに基づく、高速で使いやすい材料物性回帰モデル）
著者：Ryoma Yamamoto, Akira Takahashi, Kei Terayama, Yu Kumagai, and Fumiyasu Oba（山本竜馬、高橋亮、寺山慧、熊谷悠、大場史康）
DOI：10.1002/aisy.70592

【研究者プロフィール】

高橋 亮（タカハシ アキラ） Akira Takahashi
東京科学大学 総合研究院 フロンティア材料研究所　准教授
研究分野：マテリアルズ・インフォマティクス

大場 史康（オオバ フミヤス） Fumiyasu Oba
東京科学大学 総合研究院 フロンティア材料研究所　教授
神奈川県立産業技術総合研究所（KISTEC）　プロジェクトリーダー（兼任）
研究分野：計算材料科学

寺山 慧（テラヤマ ケイ） Kei Terayama
横浜市立大学 大学院生命医科学研究科　研究教授
研究分野：情報科学、マテリアルズ・インフォマティクス、ケモインフォマティクス